**'*WebNLG 2023 Human Evaluation*'**(AsPredicted #139263)

**Created:**       07/25/2023 12:09 AM (PT)

**Author(s)**  
Craig Thomson (University of Aberdeen) - c.thomson@abdn.ac.uk  
Anya Belz (ADAPT / Dublin City Universiy) - anya.belz@adaptcentre.ie  
  
  
**1) Have any data been collected for this study already?**  
No, no data have been collected for this study yet.  
  
**2) What's the main question being asked or hypothesis being tested in this study?**  
For each of the 4 criteria (described in the HEDS file, found in https://github.com/nlgcat/webnlg2023\_human\_eval\_preregistration), is there a statistically significant difference between the data-to-text systems evaluated, based on the judgments of human evaluators?  
  
**3) Describe the key dependent variable(s) specifying how they will be measured.**  
The dependent variable for the Fluency criterion is the participant's rating on a 1-5 scale. For all other criteria is the user response (boolean) of Yes or No. The HEDS and Repro-HEDS data sheets (which can be found via the GitHub link in #2) describe the experiment in more detail.  
  
**4) How many and which conditions will participants be assigned to?**  
Each participant is assigned 100 evaluation items, each taken from a different input item based on a Latin Square design. Each evaluation item is assigned to 1 participant.  
The number of systems submitted per language to WebNLG, and therefore the number of evaluators, is such that the number of evaluation items can be evenly divided among participants in all cases.  
Irish: 4 systems + 1 topline = 5 (20 items per system per participant)  
Maltese: 3 systems + 1 topline = 4 (25 items per system per participant)  
Russian: 3 systems + 1 topline = 4 (25 items per system per participant)  
Welsh: 3 systems + 1 topline = 4 (25 items per system per participant)  
  
**5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.**  
We will carry out four univariate ANOVAs with System as the fixed factor, and Fluency (1–5 ratings), Omission (Yes/No assessments), Addition (Yes/No assessments), and Repetition (Yes/No assessments) as the dependent variables each in one of the ANOVAs. We will report F-ratios and their statistical significance, and the homogeneous subsets of systems as determined by a post-hoc Tukey HSD analysis. The results from the latter will be reported, alongside the mean assessment values, in tables where systems whose scores are not significantly different (at the .05 level) share a letter.  
  
In addition, we will carry out factorial multivariate ANOVAs with system, evaluator and WebNLG category as the fixed factors and Fluency, Omission, Addition and Repetition as the dependent variables. We will report any significant effects of factors on dependent variables, and any significant interactions between factors. We will perform post-hoc Tukey's HSD tests to determine significant pairwise differences. We will also carry out follow-up factorial multivariate ANOVAs leaving out factors that did not display a statistically significant effect on the dependent variables.  
  
We will also report Pearson and Spearman correlations with probabilities of statistical significance, between individual Fluency, Omission, Addition and Repetition assessments.  
  
**6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.**  
All responses from participants will be kept. Validation has been enabled on the form they will be asked to complete which should prevent them from returning invalid responses. In the event that invalid responses are returned, we will ask participants to correct them.  
  
**7) How many observations will be collected or what will determine sample size?  
No need to justify decision, but be precise about exactly how the number will be determined.**  
Each language has the following number of systems being evaluated [Irish: 5, Maltese: 4, Russian: 4, Welsh: 4], with one of the "systems" being a human-authored topline in each case. There are 100 outputs evaluated per system, therefore the number of system outputs per output language is [Irish: 500, Maltese: 400, Russian: 400, Welsh: 400]. The total number of outputs is therefore 1,700.  
  
Input items are randomly sampled from the WebNLG test set, stratified for the number of triples and KB category. Then, for each of the 17 system-language pairs, a text is evaluated by 4 criteria; Fluency, Absence of Omissions, Absence of Additions, and Absence of unnecessary repetition (content). Therefore there are 1,700 judgments per criteria, for a total of 6,800 total judgments.  
  
**8) Anything else you would like to pre-register?  
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)**  
No